文書画像中のけい線・フィールドセパレータの抽出のための一手法

後藤 英昭† (正員)　　阿曽 弘具† (正員)

An Algorithm for Extraction of Rules and Field-Separators in Document image

Hideaki GOTO† and Hirotomo ASO†, *Members*

† 東北大学工学部通信工学科, 仙台市
Faculty of Engineering, Tohoku University, Sendai-shi, 980-77 Japan

あらまし　ハフ変換と自己相関関数を用いて文書画像中の任意の形状のけい線を抽出する，汎用的なけい線抽出手法を提案する．本手法により，実線はもちろん，点線や破線・鎖線，デザイン的側面の強いフィールドセパレータも文字列と分離して抽出できるようになった．

キーワード　けい線抽出，点線抽出，ハフ変換，自己相関関数

1. まえがき

文書中のけい線とフィールドセパレータには各領域の分割などを明確にする役割があり，文書画像の領域分割処理において重要な情報源となる．そのため，文書のレイアウトの抽出や理解に先立ってけい線やフィールドセパレータを抽出する処理が必要である．従来，けい線の分離・抽出のために，縦横の長さの比が大きい外接矩形をけい線とみなしたり，文書のレイアウトに関する知識を利用して囲みけい線を分離する手法が提案されている [1]．しかし，点線や破線では外接矩形が複数に分かれてしまうこと，レイアウトに関する知識を利用する場合は対象文書が限定されてしまい汎用性の向上が難しくなることが問題である．

文書画像中の外接矩形のうち，内部を占める黒画素の割合の大きいものを点線候補矩形として，その規則的な配置を手がかりとして点線を抽出する手法が提案されている [2]．この手法では，点線やそれに準じたけい線を取り扱うことができるが，実線をはじめ，より一般的な形状のけい線を統一的に取り扱うことはできない．図面の認識では画像内の線セグメントを用いて破線や鎖線を抽出する手法が用いられるが [3]，この手法では線セグメントをその配置方向に延長・結合しているため，縦縞のけい線のように線セグメントの方向が線全体の方向と異なる線や，点線は抽出できない．

本論文では，多様なけい線を抽出することを目的とした，対象けい線の種類に依存しない汎用的なけい線抽出手法を提案する．本手法は，実線・点線・破線はもちろん，線を構成する基本的な図形の形状が未知の一般的なけい線やフィールドセパレータを抽出することができ，更に，あらゆる傾きのけい線にも対応できる汎用的な手法である．本手法では，けい線(以後フィールドセパレータもけい線とみなす)の直線性の検出のためにハフ変換を用い，文字領域とけい線の区別のために，けい線の規則性に着目して，自己相関を利用している．

2. けい線抽出処理

2.1 けい線に関する知識とけい線の性質

けい線に関する一般性の高い知識として，次のようなものを利用する．

K1. 局所的に見て直線状である．

K2. 全長と幅の比が大きい．

K3. ある程度の長さをもつ．

K4. 同じ基本図形が周期的に並んでいる．(規則性)

基本図形とは，けい線を構成する基本的かつ最小の図形のことである．かすれた線や，ノイズ状で規則性を持たない線はK4を満たさないが，前者は線分の配置方向の延長よる補間処理により救済が可能と思われる．後者は実例も少なく利用される機会もほとんどないため，本論文では抽出の対象とは考えない．

文字や写真などの文書構成要素とある程度単純な形状のけい線の区別のために，次のような補助的な知識を用いる．但し，図形要素とは文字や写真を除いた文書構成要素やその一部を表し，けい線は図形要素に含まれるものとする．

AK1. 文字行に比べて幅の狭いものは図形要素である．

AK2. ある連結成分の外接矩形内部の黒画素の割合が非常に大きいものは，図形要素らしい．

AK3. 点線・破線を構成する基本図形は，それぞれある程度近い位置にある．

AK3を除き，AK1,2 はすべてのけい線に共通のものではなく，細い線と実線を効率的に抽出するのに用いる．AK1に関して，通常の文書認識では，ルビなどを考慮しても，1.4mm角(4pt)以上の大きさの文字を文字認識の対象にできればよいので，1.4mm角の文字

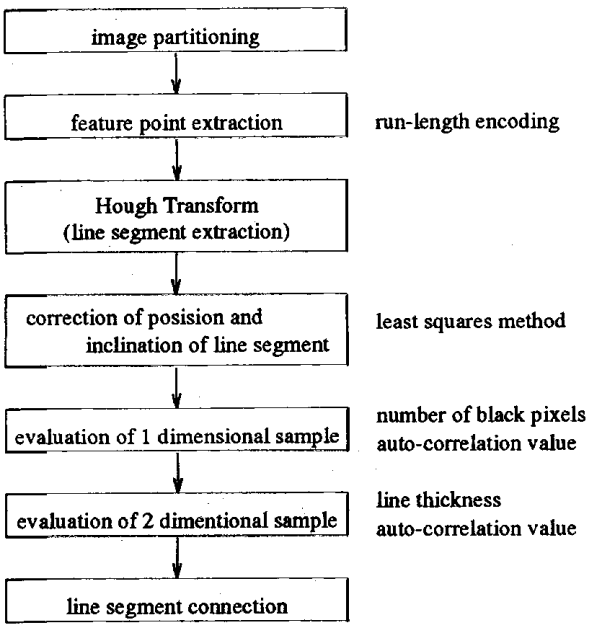

図1 けい線抽出処理
Fig. 1 Rule/Line extraction algorithm.

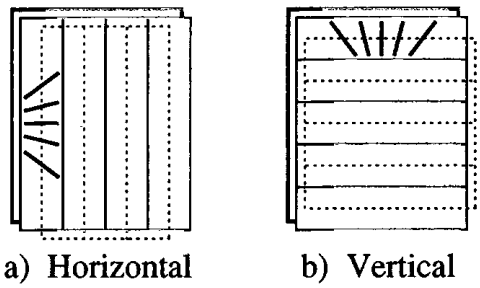

図2 画像の分割
Fig. 2 Image partitioning.

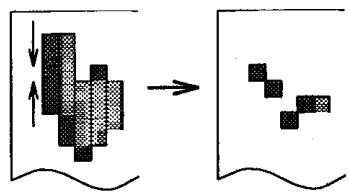

図3 黒ランの中点による特徴点
Fig. 3 Feature points by CoR.

のうち英小文字の高さを文字の最小の高さとすれば，幅が 0.7mm 未満の直線状の黒画素の集合は図形要素(けい線)とみなせる．

AK3は，点線などのドット間の空白の長さとドットの長さの比(黒画素の割合)に関する知識を表す．様々な文書の中のけい線について調査したところ，点線に限らず，大多数のけい線は芯線付近の黒画素の割合が1/4以上のものであった．本研究ではこのことを仮定して，けい線の候補を絞り込むことにした．

2.2 画像の分割

図1に，けい線抽出処理の流れを示す．線素の結合以降の処理には従来の手法[3]が利用できるので，以下，これより前の各処理の内容について説明する．

処理の高速化のために，まず元の文書画像(通常400 ～ 600dpi)から再量子化により200dpi(約8ドット/mm)の解像度の画像を作る．画像を図2に示すように横方向と縦方向に分割し，それぞれを独立に処理対象とする．画像を分割してできる長方形の小領域をバンド，その幅 W をバンド幅と呼ぶ．境界の影響を抑えるために，隣合うバンドは互いにバンド幅の半分だけ重なるようにする．バンド内で検出される線分を線素と呼ぶ．それぞれの分割方向において $[-45°, 45°)$, $[45°, 135°)$ の角度の範囲で線素を抽出する．実際の文書で用いられているけい線の太さと長さ，基本図形の繰返し周期を考慮して，バンド幅を $W = 64$ ドット(=8mm, 200dpi)，知識K3における「ある程度の長さ」をバンド幅の3/4(約6mm)に設定した．

2.3 特徴点の抽出とハフ変換

ハフ変換の処理の高速化のためには，特徴点を少なくすることが有効である．画像をランレングス符号化した時の各黒ランの中点をCoR(Center of Run)と呼び，これを特徴点として用いることにする．図3に，画像を横方向に分割した場合の特徴点の生成の様子を示す．傾きが $[-45°, 45°)$ の線素を抽出する場合は縦方向，$[45°, 135°)$ の場合は横方向のランレングス符号を用いる．CoRを2次元平面上にプロットして得られる画像をドットイメージと呼ぶ．この他，細線化された画像の黒画素を特徴点とする手法についても検討する．

線素の検出のためにハフ変換を使う．ここで用いる直線検出用のハフ変換は，直線の式を $y = ax + b$ で表し，a, b をパラメータとするものである．処理の高速化のために，黒画素を優先した再量子化により画像の横方向(分割の方向)の解像度を100dpiに落とし，これから作成したCoRを特徴点として用いた．解像度200dpiの文書画像のサンプリング間隔を $\Delta s = 1$ として，パラメータ b, a のサンプリング間隔をそれぞれ $\Delta b = 2 \cdot \Delta s = 2$, $\Delta a = \Delta b/(W/2) = 1/16$ とした．

パラメータ空間中のピークの検出に先立ち，ピークの検出精度の向上のために，パラメータ a の各標本点において b 軸に並行な方向に，側抑性のある1次元的なフィルタをかける．ここで用いるフィルタは，要素

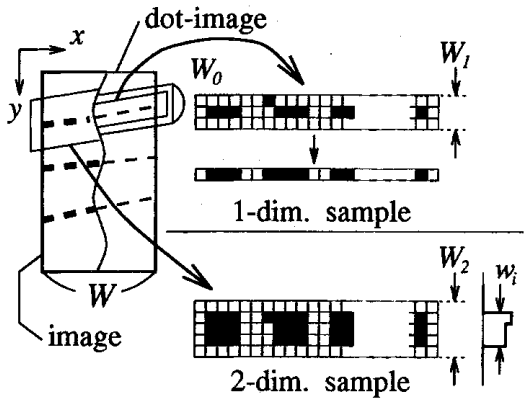

図4 1次元，2次元サンプル
Fig. 4 1- and 2-dimensional samples.

が $[-1, 1/2, 1, 1/2, -1]$ で表されるFIR型フィルタである．続いてパラメータ空間からピーク値がしきい値 T_p 以上のピーク点(極大点)を抽出する．T_p の値は，抽出すべき線素に対するピーク値の最低値(予測値)に設定する．本研究では予備実験により $T_p = 5$ とした．i 番目のピークにおいて，ピーク値を $W/2$ で割って正規化した値を h_i とおく．h_i は線素の芯線付近の黒画素数の割合に相当する．

2.4 1次元的なサンプルの評価

ハフ変換によって得られたピーク点のパラメータ値に対応する線分を，線素候補と呼ぶ．はじめに，解像度200dpiの画像のCoRから得られたドットイメージにおいて，各線素候補を中心とした幅 W_0 の部分画像に注目し，この中にあるCoRを用いて，最小2乗法により線素候補の位置 b と傾き a の値を補正する．補正後の線素候補の周りに，幅 W_1 の部分画像を切り出す．(図4右上) この部分画像の垂直方向の射影分布を計数値のゼロと非ゼロで2値化して，長さが W ドットの1次元サンプル(1次元ビット列)を得る．部分画像に幅をもたせているのは，特徴点の位置のずれを吸収するためであり，実験では $W_0 = 7$, $W_1 = 5$(ドット)に設定している．知識K3より，1次元サンプルの両端の空白を除いた長さが $3W/4$ に満たなかったり，長い空白 $(> 3W/4)$ を含む場合は，対応する線素候補を破棄する．

1次元サンプル内の黒画素の総数がしきい値 T_{DC} 以上であれば，対応する線素を実線として取り扱う．T_{DC} は，実線がかすれにより1割程度の黒画素を失っても，それを実線として取り扱えるような値に決める．$T_{DC} = (3W/4) \times 0.9 \simeq 43$(ドット) とした．

黒画素数が T_{DC} に満たないものについては，自己相関関数を計算してけい線らしさを評価する．白画素を -1，黒画素を1に対応させて，i 番目の線素候補に対応する1次元サンプルを $s_i(n) \in \{-1, 1\}, (0 \leq n < W)$ と表す．自己相関関数 $M_i(l)$ は，

$$M_i(l) = \frac{1}{W} \sum_{n=0}^{W-1} s_i(n) s_i(n+l) \qquad (1)$$

と表される．但し，$n \geqq W$ に対して周期的な境界条件 $s_i(n) = s_i(n - W)$ を用いる．自己相関関数の $l = 0$ に最も近い正の l における第1ピークの高さを自己相関値と呼び，m_i と表す．第1ピークが $0 \leqq l \leqq W/3$ の範囲になかったり，自己相関値がしきい値 T_L に満たない場合は，基本図形の規則性が不十分とみなして対応する線素候補を破棄する．

ハフ変換で得られたピーク値 h_i や自己相関値 m_i が大きいほど，その線素はけい線らしいと言える．自己相関値 m_i は線素候補をずらして重ね合わせたときに一致する画素数の割合の極大値に相当するので，h_i, m_i はいずれも画素数を単位とする．本研究では，しきい値を T_C とおき，次の条件を満たす線素だけを次段の処理に引き渡すことにした．

$$h_i + m_i > T_C \qquad (2)$$

2.5 2次元的なサンプルの評価

次に，2次元的なサンプルを用いた詳細な識別を行う．図4右下のように，線素に平行な方向の部分画像を2次元サンプルとして切り出し，射影分布を用いて線素の太さ w_i を求める．本研究では2次元サンプルの高さを $W_2 = 33$ ドット(約4mm)に制限しているが，これは太いけい線に対しても十分な値である．

知識AK1から，$w_i \leqq 5.5$ ドット(0.7mm) となる線素はけい線とみなすことができる．しかし本けい線抽出手法ではあらかじめ1次元サンプルを用いてけい線らしさがある程度評価されているので，これよりも幅の広い線素候補をけい線とみなしても誤抽出はほとんど増加しない．そこで，2次元サンプルの自己相関関数の計算を減らして処理を高速化するために，$w_i \leqq 16$ ドットとなる線素候補をけい線とみなす．

残りの線素については，$W \times w_i$ の部分画像における自己相関関数を求める．この自己相関関数は，まず線素に平行な方向の各ラスタごとの自己相関関数を求めた後で，ラスタ間での平均をとったものである．1次元の場合と同様に自己相関値を求め，その値がしき

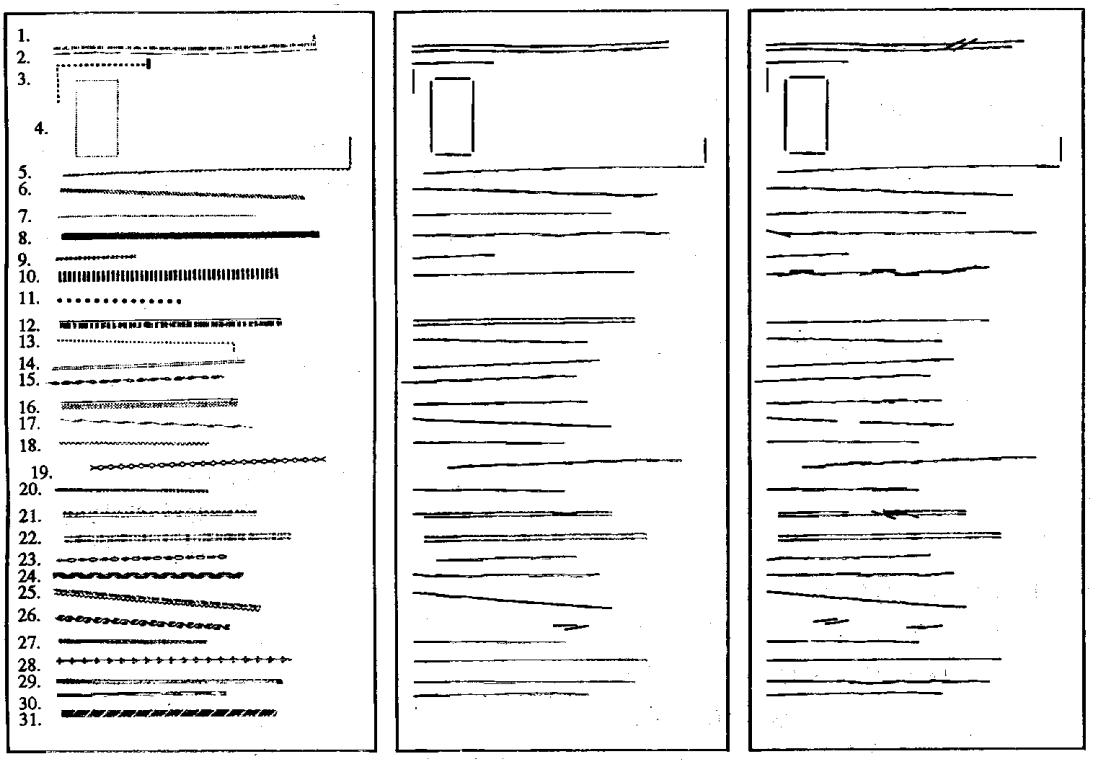

図5 けい線抽出処理の例
Fig. 5 An example of rule/line extraction.

い値 T_A を超える場合に対応する線素をけい線とみなす．また知識AK2より，部分画像内の黒画素の割合が大きい(ここでは80%以上とした)線素もけい線とみなす．

3. 実験・結果

始めに，新聞などから無作為に集めた図5 a)のNo.1～10, 12～16のけい線を用いて，各パラメータを調整した．パラメータ T_L, T_C の値は，No.1～16 (No.11を除く)のけい線から生じた線素がすべてけい線と判断されるように定めた．パラメータ T_A の値は，上記と同じけい線の他に文字を大量に含む文書も用いて，なるべくけい線の取りこぼしがなくなるような方針で最適値を求めた．特徴点に黒ランの中点(CoR)を用いた場合の各パラメータの値はそれぞれ，$T_L = 0.3, T_C = 0.65, T_A = 0.63$ である．

図5 b),c) にけい線の抽出結果をけい線の中心線で表した．b)はCoRによる特徴点，c)は細線化画像による特徴点を用いた結果である．細線化のアルゴリズムは十近傍法[4]を用いた．b) では No.11,26,31 を除いて，良好な結果を得ている．抽出に失敗したけい線はいずれも基本図形の繰り返しの周期が長いものである．細線化画像を用いた場合は線素のとりこぼしが多くなり，No.17,21 のようにけい線の分断が見られる．点線に細線化が施された場合，点線の長手方向にもドットが縮み，黒画素の割合が小さくなったことがとりこぼしの原因である．この結果から，細線化を用いることより，けい線の方向をある程度仮定して作られるCoRによる画像表現の方がけい線の抽出には適切であると判断される．

文字を含む画像に対するけい線抽出処理の評価のためにTIME誌21ページ(注1)を処理したところ，けい線はすべて正しく抽出できた．文字をけい線とみなす誤認識が1ページあたり5, 6箇所に見られたが，ほとんどのけい線候補の長さは8mm程度であり(最大30mm程度)，文字領域の面積に占める割合は非常に小さい．誤抽出の中には，文字の並びがけい線に酷似しているなど，やむを得ないと思われるものが少なくない．これらを正しく取り除くためには，文字の大き

(注1)："TIME JANUARY 11", Vol.141, No.2, pp.8–13, 15–29(1993).

さや文書構造に関する知識を利用する必要がある．

特徴点にCoRを用いた場合の本けい線抽出手法の処理時間は，文字を含むB5版の文書画像(本誌表紙)において全体で112秒(SPARCstation10/M51)であった．内訳は処理時間の長い順に，線素の位置・角度の補正が30秒，2次元サンプルの評価が25秒，1次元サンプルの評価が21秒，その他である．抽出できるけい線を水平方向 $-7^\circ \sim 7^\circ$ と垂直方向 $83^\circ \sim 97^\circ$ に制限した場合は，全体の処理時間は39秒となった．

4. むすび

ハフ変換と自己相関関数を用いて，文書画像から任意の形状のけい線を分離・抽出する手法を提案した．パラメータの調整では用いなかった未知文書を含めた実験により，幅の狭いけい線については提案手法が高いけい線抽出能力をもつことが示された．

太いけい線や基本図形の周期の長いフィールドセパレータに対しても安定な抽出を可能とすること，処理の効率化，文字領域の誤認識の低減などが今後の課題である．

文　献

[1] 岡本正行, 宮澤　章, "多様なレイアウト構造, 構成要素を持つ文書の領域分割," 信学技報, PRU89-114, Feb. 1990.

[2] 園田浩一郎, 米田政明, 長谷博行, 酒井　充, "文書画像中の点線抽出アルゴリズムの検討," 信学'94秋全, D-316.

[3] 嶋田　茂, 角本　繁, 江尻正員, "図面自動入力のための破線・鎖線の認識アルゴリズム," 信学論(D), vol.J69-D, no.5, pp.759-770, May 1986.

[4] 阿曽弘具, "並列処理向き細線化アルゴリズム," 信学論(D-II), vol.J76-D-II, no.9, pp.2148-2150, Sept. 1993.

(平成7年4月3日受付，8月8日再受付)